11-3-2017

Launching a Web Archives Program at a Public University

Blake Graham
University of Nebraska-Lincoln, blake.graham@unl.edu

Jennifer L. Thoegersen
University of Nebraska-Lincoln, Jennifer.Thoegersen@oslomet.no

Mary Ellen Ducey
University of Nebraska - Lincoln, mducey2@unl.edu

Launching a Web Archives Program at a Public University

Blake Graham
Digital Archivist
University of Nebraska-Lincoln

Jennifer L. Thoegersen
Data Curation Librarian
University of Nebraska-Lincoln

Mary Ellen Ducey
University Archivist/Special Collections Librarian
University of Nebraska-Lincoln

Abstract

Many organizations and institutions rely heavily on a web presence to disseminate information and to manage programs and policies. This tendency leaves library and archive professionals with a challenge: how best to capture and preserve web-based information and resources. Over the last few years, the proactive collection and management of web archives has gained traction across all types of libraries and archival repositories. This paper offers a synopsis of actions and initiatives conducted by a small team dedicated to creating a sustainable web archives program at the University of Nebraska-Lincoln Libraries. The authors discuss (1) how the project team formed and the complementary skill sets of the group; (2) the details of the project, including the project scope, objectives, and timeline; (3) the identification and selection process for web resources; (4) the approach for testing and implementing a web capture tool, using Archive-It as an example; and (5) ongoing efforts and challenges for web archives at the university. The write-up is geared towards a broad audience of information professionals in cultural heritage institutions that are interested in project management in libraries and web archives in general.

Introduction

On January 31, 1997, the Internet Archive first archived the main page of the University of Nebraska-Lincoln's (UNL) website (http://www.unl.edu) (Internet Archive, 2017). As a non-profit founded in 1996, the Internet Archive is undertaking what is likely the most well-known and ambitious project to save and provide access to websites as they appeared over time (Internet Archive, n.d.). Given the vastness of the World Wide Web, the Internet Archive takes a broad yet limited approach by archiving many websites, but not always frequently or deeply into the website's structure. As such, cultural heritage institutions have increasingly undertaken more curated archiving of websites that are important to their missions and the communities they serve.

UNL Libraries' interest in beginning a web archives program goes back several years. Thanks to the work of many individuals within the organization, the Libraries are now a member of

Archive-It, the Internet Archive's subscription web archiving service, allowing the institution to identify, archive, and preserve web content of historic importance to the university.

Review of Literature

Building a web archives program presents several challenges for any institution. Due to the changing nature of the web as well as the volume and complexity of digital objects, the process of web archiving requires measures beyond the scope of traditional archival practices and workflows. The National Library of Australia (NLA), for example, shared their difficulty in making preservation plans and decisions for web archives. These difficulties were due to (1) the inability to help guide or otherwise control the creation of original content and corresponding format, standards, or quality; (2) the methodological deficiency in collecting and rendering web archives; and (3) the inherent flaws of taking time-limited "snapshots" of dynamic content (Webb, Pearson, & Koerbin, 2013). These challenges pose a new threat to Libraries, Archives, and Museums (LAMs), and require meaningful discussion and iterative, creative preservation planning.

Staffing is another area of challenge expressed in several national and international surveys. In 2016, the National Digital Stewardship Alliance (NDSA) conducted a survey on web archiving in the United States. The resultant report showed that among the institutions that have web archiving initiatives (n=84), only 24% had one or more FTE dedicated to web archiving tasks and more than half devoted only 0.25 FTE (Bailey, Grotke, McCain, Moffatt, & Taylor, 2017). These numbers were similar to those presented in NDSA's 2013 report, which suggested that "research, development, and technical experimentation necessary to advance the archiving tools on these fronts will not come from the majority of web archiving organizations with their fractional staff time commitments" (Bailey et al., 2014, p. 22). In 2015, a similar survey by Harvard Library reached the same conclusion – namely, that the majority of institutions from across the world with established web archiving programs have no full-time staff dedicated to web archive projects (Truman, 2016). These reports suggest that as organizations increasingly invest in web archiving activities and initiatives, allocation for manpower is not increasing proportionally, which can often exhaust operational capabilities at the local level.

Lastly, variations of metadata application appear to be a recurring topic for discussions on planning and implementation of web archives. In response to a 2015 OCLC partner survey, OCLC Research created a Web Archiving Metadata Working Group (WAMWG) to address both metadata guidelines and the use of web archives (Erway, 2015). One of the early efforts of the working group included collecting and analyzing local documentation on metadata application. Seven guidelines from different organizations were compared, and the working group confirmed a lack of shared practices for metadata application throughout the professional community (Dooley & Bowers, in press; Dooley, Farrell, Kim, & Venlet, 2017). Different institutions are applying common metadata standards in different ways. For example, WAMWG's findings reveal that the *Date* field could potentially be expressed by the copyright date within the website, the date of capture within Archive-It, the beginning and end dates of the site's existence, or the origination date of the content displayed within the site (Dooley & Bowers, in press). In practice, non-uniform application of descriptive metadata standards will breed inconsistencies and (eventually) compatibility issues. As WAMWG describes it, "the need for sustainable practices,

in light of limited staff resources, poses an enormous challenge for metadata creation" (Dooley & Bowers, in press, p. 5).

While challenges in decision-making processes, staffing, and metadata application often complicate implementation, there are other, counteractive trends surfacing across LAMs as well. Local and regional collaboration and partnerships are forming to create efficiencies in several areas, including documentation during (and after) implementation. The successful consolidation of documentation and budgetary considerations by the Kansas Archive-It Consortium (KAIC) serves as an excellent example of overcoming such hurdles (Hight, Todd-Diaz, Schulte, & Church, 2017). Similarly, with support from the Andrew W. Mellon Foundation, the New York Art Resources Consortium successfully initiated a collaborative program of web archiving focused on specialist art historical resources (Duncan, 2016). One of the products of this grant included a public metadata application profile for online art resources, such as auction catalogues, catalogues raisonnés, and artists' websites (Guenther, 2015). The Mellon Foundation also awarded Columbia University Libraries a similar grant for 2013-2015 with the explicit goal of fostering web archiving collaboration via the Web Resources Collection Program (Columbia University Libraries, 2017). Lastly, the California Digital Library, which is one of the most-successful collaborative web archiving programs in the country, has joined together 11 library systems to expand collective capacity to steward web archive collections (California Digital Library, n.d.).

The idea and practice of collaboration for web archives also extends beyond the scope of institution-driven projects. Many archivists and curators are partnering with a wide variety of users and community groups to help build web archives. Sylvie Rollason-Cass and Scott Reid (2015) discuss the profound progress that can be accomplished through institutional collaboration and community-based partnerships, especially web archives on social movements. There's growing evidence to suggest that although recent surveys have shown little to no growth in staffing models across recently-launched and established web archives programs, there is an abundance of community-based working groups, partnerships, and inter-institutional collaborations forming to spearhead the many challenges on the horizon.

Project History & Planning

UNL Libraries Archives & Special Collections (UNL Archives) attempted to build and integrate a web archives program to adapt to the changing nature of publication methods at UNL. The effort to harvest, preserve, and provide access to websites created by the university is in accordance with policy outlined by the Board of Regents. Specifically, the university's Records Retention Policy dictates that UNL "has a responsibility to preserve the history of the University for future generations" (UNL Business & Finance, n.d.).

Early efforts to address web archiving at UNL did not have the benefit of formal funding, staffing, or policies. The first major steps toward a web archiving program occurred in 2014 with the implementation of the Rosetta preservation system. Initial plans included use of the open source Web Curator Tool (WCT) software (which has an integration with Rosetta) to manage web archiving. Even with the well-documented WCT software and extremely supportive software developers, UNL Archives struggled to get the software running properly in a

production environment at the time. In late 2014, the Libraries prioritized using Rosetta to ingest current digital content rather than implementing web archiving, and put work on the WCT on hold.

In early 2015, a committee consisting of the data curation librarian, archivists from UNL Archives, and a library science graduate student developed a software-agnostic project to lay the groundwork for a web archiving program. Models developed by the University of Michigan and Indiana University guided the development of the project charter (Shallcross, 2011; Indiana University, n.d.). The charter focused on three main areas: (1) an inventory of domains and subdomains targeted for archiving, (2) the application of descriptive metadata standards, and (3) a review of staff and funding required to implement and sustain a web archiving program. Milestones for the project included identification of websites, guidelines on archiving frequency and depth, archiving non-HTML file formats, using specific standards, addressing intellectual property and copyright concerns, training, and access and delivery tools.

While retention guidelines from the university provided guidance on what web content to archive, UNL Archives focused generally on archiving three types of records: (1) administrative records that are kept for legal, financial, or long-term historical purposes; (2) faculty records where individuals own intellectual property and copyrights; and (3) recognized student organizations, owned by students as developers of organizational property. Using this as a starting point, the initial round of discovery revealed over 325 websites that would be candidates for inclusion. Specific sites were selected as examples of important and varied websites, including sites for administrative units, such as the Office of the Chancellor and academic departments, and websites of the broader University of Nebraska (NU) system. Other entities included the NU system Board of Regents, the Office of Vice Chancellor for Research and Economic Development, Student Affairs, and the Institute of Agriculture and Natural Resources. The review outlined the priority level for each site, its provenance, how frequently it should be archived, and considerations of providing additional metadata for access. This index of potential sites to archive was later repurposed during the implementation phase, mentioned below.

During the planning phase, UNL Archives grew interested in Archive-It. As a service of the Internet Archive, Archive-It provides an easy-to-use platform to manage web archiving activities. Many of UNL's fellow Big Ten Academic Alliance institutions were already using Archive-It and seemed pleased with the pricing and results. Whether UNL selected Archive-It or another tool, UNL Archives determined that additional resources would be necessary to implement a web archives program, including funds to pay for the selected tool and staff time to implement the program and manage it over time.

The web archiving project gained significant momentum in late 2016 when UNL Libraries hired a digital archivist. Among other responsibilities, this position was largely responsible for leading the web archiving program. Following a trial of the service, a new project to implement Archive-It received approval from the Libraries' administration on February 1, 2017. Shortly thereafter, in spring 2017, UNL Archives officially adopted Archive-It as an instrument for crawling, managing, and providing access to web data created by campus entities.

Implementation

Training and practical application drove the earliest stage of implementation. Using Archive-It's documentation and training videos, UNL Archives began learning about the features of crawling and archiving websites and started establishing broad groupings of sites, or "collections." Each collection in Archive-It represents a primary collecting area for Archives and Special Collections, traditionally known as either a record group or manuscript collection. In most cases, each collection in Archive-It represents an existing physical collection (UNL Libraries, Archives and Special Collections, n.d.). Sites, or "seeds," related to the Vice Chancellor of Business and Finance, for example, would be assigned to RG28 – the designated record group for UNL's Business and Finance records (see Figure 1 for Archive-It's account structure). As a result, a total of 30 primary collecting areas were initially created in Archive-It. Each collection has six metadata fields to represent the nature and scope of the collecting area (see Figure 1).

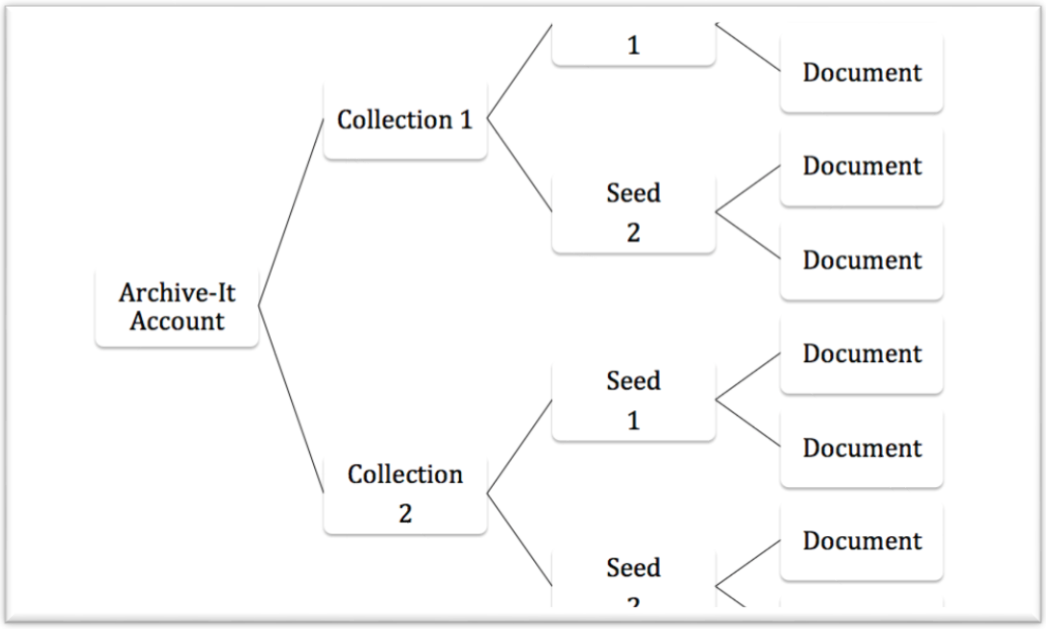


Figure 1. Archive-It's Account Structure for Collections (Praetzellis, 2017a).

Figure 2. Screenshot of Metadata Fields for Record Group 18.

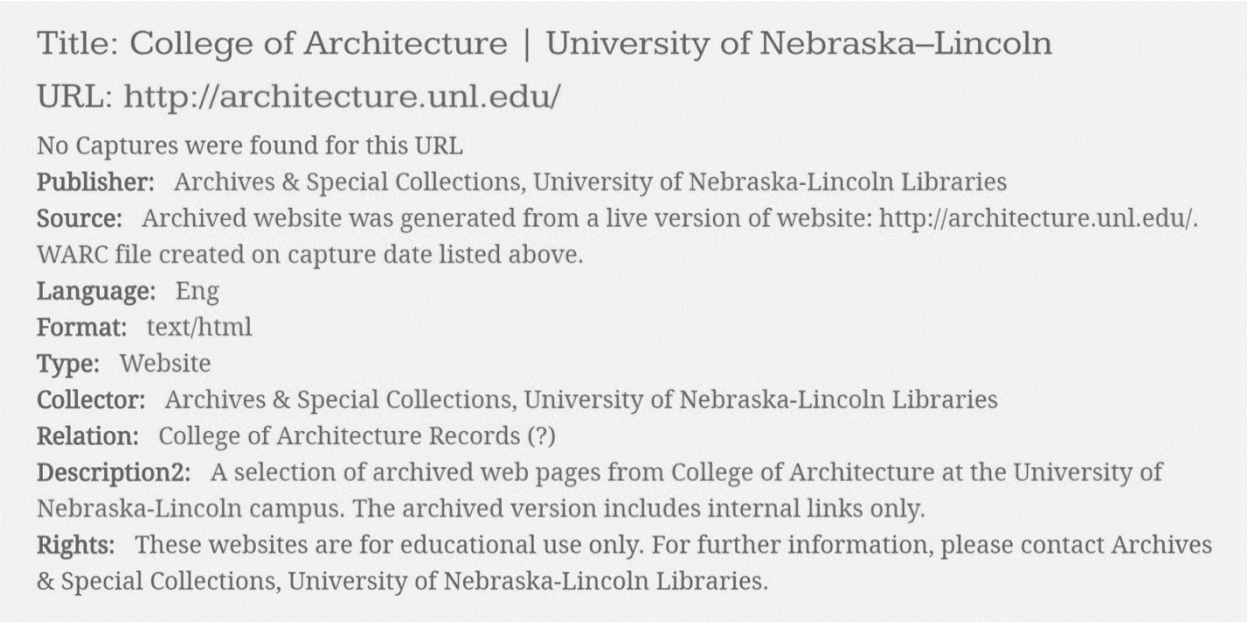
While setting the primary collecting areas for web archives collections, UNL Archives began repurposing an existing index of potential sites and forming a master spreadsheet of seeds. The columns in the spreadsheet were expanded to include: priority, collection, title, contributor, URL, source, and finding aid URL. In particular, priority, URL, and finding aid URL were the three fields needed for appraisal-related actions. Since each Archive-It member has a data limit, setting a priority for each seed enabled the team to sort seeds and impose data limits on lower-priority sites.

Sites were continually added to the master sheet using two methods: manually browsing the site structure of the www.unl.edu domain, and crawling and exporting a sitemap. The latter proved to be the more efficient method, but the sitemap generator began slowing down while attempting to create long indexes of inconsequential subpages. For example, if a single subdomain (e.g., http://snr.unl.edu/) contained thousands of subdirectories, then a sitemap generator would attempt to find every page structurally related to the subdomain. Unfortunately, many sitemap generators are truncated when the list exceeds several thousand pages. And, since UNL Libraries crawls primary subdomains and other valuable sites for UNL, a comprehensive list of all subdirectories was unnecessary. A small portion of subdomains were retained from the sitemap generator, but most new additions to the master list were created by manually navigating the www.unl.edu domain.

Once the master list was completed, the team copied and pasted URL's from the spreadsheet into a corresponding collecting area in Archive-It. Roughly 175 sites were carefully reviewed and placed within a designated collection. After transplanting the seeds, the team created seed-level metadata in bulk using the “Bulk Seed Metadata” feature. The team simply downloaded the csv template, and began populating fields using Microsoft Excel and OpenRefine.

In order to begin creating original metadata for seeds, the department chose to utilize two data dictionaries created by the New York Art Resources Consortium (NYARC) and the WAMWG as guides during the process (Guenther, 2015; Dooley & Bowers, 2017). The former guide is primarily based on MARC fields, with mappings to Dublin Core fields, while the latter is a

hybrid set of elements based on DACS, RDA, and Dublin Core. In a timely coincidence, the WAMWG released a preliminary draft of their data dictionary while the UNL Archives team was developing metadata for seeds. Using both as guides, the team attempted to create Dublin Core-based metadata records for most of the seeds, and upload the "bulk seed metadata" csv file into Archive-It (see Figure 3). It's worth noting that both Subject and Dates fields were largely ignored during this process. Subjects were postponed until after the implementation phase, and Dates were automatically populated in Archive-It during crawl sessions.


Title: College of Architecture | University of Nebraska–Lincoln

URL: http://architecture.unl.edu/

No Captures were found for this URL

Publisher: Archives & Special Collections, University of Nebraska-Lincoln Libraries

Source: Archived website was generated from a live version of website: http://architecture.unl.edu/. WARC file created on capture date listed above.

Language: Eng

Format: text/html

Type: Website

Collector: Archives & Special Collections, University of Nebraska-Lincoln Libraries

Relation: College of Architecture Records (?)

Description2: A selection of archived web pages from College of Architecture at the University of Nebraska-Lincoln campus. The archived version includes internal links only.

Rights: These websites are for educational use only. For further information, please contact Archives & Special Collections, University of Nebraska-Lincoln Libraries.

Figure 3. Screenshot View of Seed-Level Metadata in a Collection

After the seeds and collections were organized, the team began crawling each seed. Crawling – an operation in Archive-It that identifies materials on the live web to become archived content – and monitoring crawl results are recursive processes for any web archives program. Archive-It provides training videos, as well as a support ticket system, to better understand and troubleshoot crawling procedures, such as modifying crawl scope and data limits, bypassing robots.txt, and avoiding crawler traps (Praetzellis, 2017b). Most crawls are fairly straightforward; members enter a seed URL (e.g., https://nebraska.edu/publications-and-reports/), and the Heritrix web crawler then identifies and copies live content.

Once a crawl is completed, Archive-It generates a report to check crawl results and determine whether to perform a "patch crawl" if necessary to retrieve any missing content. As an example, a patch crawl would be needed if an institution sets the crawl scope to "standard" in Archive-It, and the site being crawled contains external links to other websites (e.g., YouTube videos); in this case, those external websites would be considered out of scope. If the external links are deemed important enough to include in the archived content, then crawling features can be enabled so the archived content can be patched and specific external links can be included. This crawl-and-report method proved incredibly valuable for the team, because systematically crawling all websites and external links would clutter the collections.

At the conclusion of the implementation phase, the UNL Archives team will shift focus to drafting metadata application guidelines, assigning subject headings for seeds, and conducting patch-crawling work for incomplete crawls. Beyond this work, the team is mindful of the fact that this new responsibility requires proactively monitoring web archives collections over time, which adds a new component to the department's existing responsibilities and services. While Archive-It will remain the primary tool for crawling, archiving, and providing access to archived websites, Rosetta will serve as the preservation system holding a copy of the Archive-It data. The WARC files generated in Archive-It will be downloaded locally on an annual basis at the end of the fiscal year, and then moved into a preservation environment.

Conclusion

One aspect of implementation that extends beyond practical application is integrating web archiving practices into traditional archival workflows. At UNL Libraries Archives & Special Collections, integrating web archives required a deep knowledge of existing skills among staff, tools used in different work areas, documentation and local practices, and the various conditions and processes that impact the workflow of creating and managing digital collections. As with many repositories, areas of the department are transitioning to meet the changing needs of both users and collections management.

Collaboration between individuals and departments in the UNL Libraries made the creation of a web archives program and the implementation of Archive-It possible. As the program continues to gain momentum, UNL Archives hopes to expand collaboration to other institutions, especially with colleagues at the three other NU system campuses.

Proper funding and staffing are essential for a successful web archives program. However, the initial lack of funding and staffing encouraged UNL Archives to focus on planning and laying the groundwork for future work. This early initiative ensured the team was prepared to implement the web archives program once staff and funding became available. Now, the adoption and implementation of Archive-It provides a method of archiving the web-based output of the university over time. Being able to preserve and provide access to this content has ensured UNL Archives is better able to fulfill its mission both now and in the years to come.

References

Bailey, J., Grotke, A., Hanna, K., Hartman, C., McCain, E., Moffatt, C., & Taylor, N. (2014). *Web archiving in the United States: A 2013 survey*. Retrieved from http://ndsa.org/publications/

Bailey, J., Grotke, A., McCain, E., Moffatt, C., & Taylor, N. (2017). *Web archiving in the United States: A 2016 survey*. Retrieved from http://ndsa.org/documents/WebArchivingintheUnitedStates_A2016Survey.pdf

California Digital Library. (n.d.). *Web archiving activities* [Project website]. University of California Curation Center. Retrieved from http://www.cdlib.org/services/uc3/webarchiving/

Columbia University Libraries. (2017). *Web resources collection program* [Project site]. Retrieved from https://library.columbia.edu/bts/web_resources_collection.html

Costa, M., Gomes, D., & Silva, M. J. (2016). The evolution of web archiving. *International Journal on Digital Libraries*, 1-15. doi: 10.1007/s00799-016-0171-9

Dooley, J. M., Farrell, K. S., Kim, T., & Venlet, J. (2017). Developing web archiving metadata best practices to meet user needs. *Journal of Western Archives, 8*(2), 1-14. Retrieved from http://digitalcommons.usu.edu/westernarchives/vol8/iss2/5

Dooley, J., & Bowers, K. (in press). *Best practices for web archiving metadata: Report of the OCLC Research Library Partnership* [draft]. Retrieved from https://drive.google.com/file/d/0B7XD89gl28x5dlhZVUE2VFlDdms/view

Dublin Core Metadata Initiative. (n.d.). Glossary: One-to-one principle. Retrieved June 9, 2017 from the DCMI Wiki: http://wiki.dublincore.org/index.php/Glossary/One-to-One_Principle

Duncan, S. (2016). Web archiving at the New York Art Resources Consortium (NYARC) [Digital Library Federation blog]. Retrieved from https://www.diglib.org/archives/11227/

Erway, R. (2015). Thoughts from partner staff about web archiving [OCLC Research blog]. Retrieved from http://hangingtogether.org/?p=5450

Guenther, R. (2015, June 1). *Metadata application profile and data dictionary for description of websites with archived versions*. Retrieved from http://www.nyarc.org/press

Hight, C., Todd-Diaz, A., Schulte, R., & Church, M. (2017). Collaboration made it happen! The Kansas Archive-It Consortium. *Journal of Western Archives, 8*(2), 1-25. Retrieved from http://digitalcommons.usu.edu/westernarchives/vol8/iss2/4

Indiana University. (n.d.). *What is the Archives of Institutional Memory (AIM)?* Indiana University. Retrieved from http://institutionalmemory.iu.edu/aim/

Internet Archive WayBackMachine. (2017). *http:/www.unl.edu:80/*. Retrieved from https://web.archive.org/web/19970801000000*/http://www.unl.edu:80/

Internet Archive. (n.d.) *About the Internet Archive*. Retrieved from https://archive.org/about/

Miller, S. J. (2011). *Metadata for digital collections: a how-to-do-it manual*. New York, NY: Neal-Schuman Publishers.

Praetzellis, M. (2017a). Archive-It trial basics [Help center website]. Retrieved from https://support.archive-it.org/hc/en-us/articles/208111766-Archive-It-Trial-Basics

Praetzellis, M. (2017b). Glossary of Archive-It and web archiving terms [Help center website]. Retrieved from https://support.archive-it.org/hc/en-us/articles/208111686-Glossary-of-Archive-It-and-Web-Archiving-Terms

Rollason-Cass, S., & Reed, S. (2015). Living movements, living archives: Selecting and archiving web content during times of social unrest. *New Review of Information Networking, 20*(1-2), 241-247.

Shallcross, M. (2011, April). On the development of the University of Michigan Web Archives: Archival principles and strategies. In *Society of American Archivists Campus Case Studies*. Retrieved from http://files.archivists.org/pubs/CampusCaseStudies/Case13Final.pdf

Truman, G. (2016). *Web archiving environmental scan: Harvard Library report*. Retrieved from http://nrs.harvard.edu/urn-3:HUL.InstRepos:25658314

UNL Business & Finance. (n.d.) *Records retention policy*. Retrieved from http://bf.unl.edu/policies/bf/RecordsRetention.shtml

UNL Libraries, Archives and Special Collections. (n.d.) *Mission and collection scope*. Retrieved from http://libraries.unl.edu/archives-special-collections-mission-collection-scope

Webb, C., Pearson, D., & Koerbin, P. (2013). 'Oh, you wanted us to preserve that?!' Statements of preservation intent for the National Library of Australia's digital collections. *D-Lib Magazine, 19*(1/2), 2.